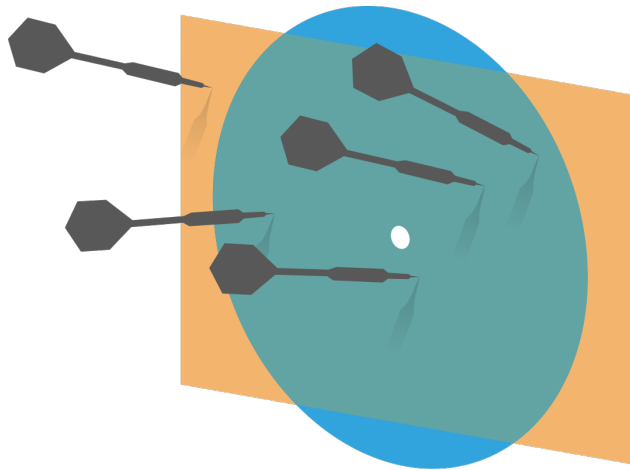


Probability and Statistics for Computer Science



Credit: wikipedia

“Unsupervised learning is arguably more typical of human and animal learning...”--- Kelvin Murphy, former professor at UBC

Last time

- ✱ Curse of dimensions
- ✱ Unsupervised learning
- ✱ Clustering

Q. Is k-means clustering deterministic?

A. Yes

B. No

Some issues with k-means clustering

- ✱ Sensitive to outlier
- ✱ Sensitive to the seeds (centroids)

Some issues with k-means clustering

- ✱ Sensitive to outlier: example

Some issues with k-means clustering

- ✱ Sensitive to the seeds (example)

K-means clustering example: Portugal consumers

- ✱ The dataset consists of the annual grocery spending of 440 customers
- ✱ Each customer's spending is recorded in 6 features:
 - ✱ fresh food, milk, grocery, frozen, detergents/paper, delicatessen
- ✱ Each customer is labeled by: 6 labels in total
 - ✱ Channel (Channel 1 & 2) (Horeca 298, Retail 142)
 - ✱ Region (Region 1, 2 & 3) (Lisbon 77, Oporto 47, Other 316)

Lisbon, Portugal

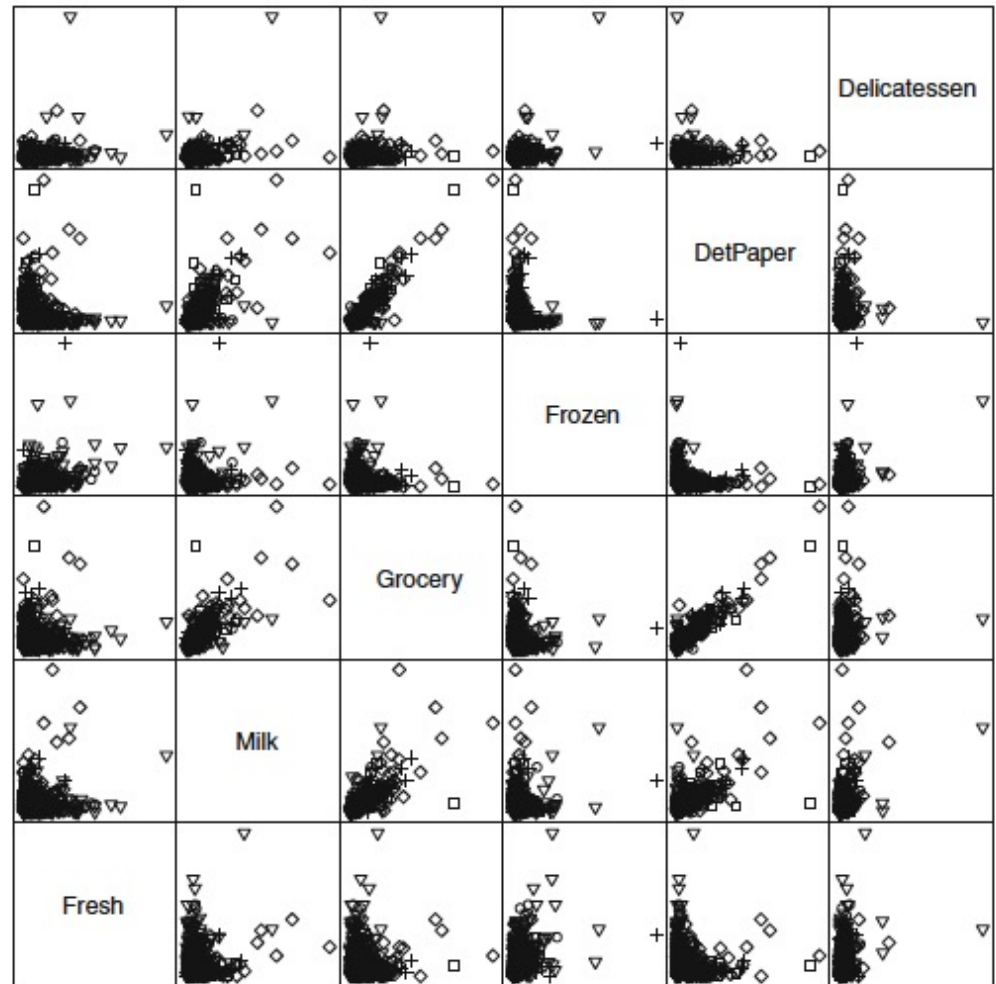


Oporto, Portugal



Visualization of the data

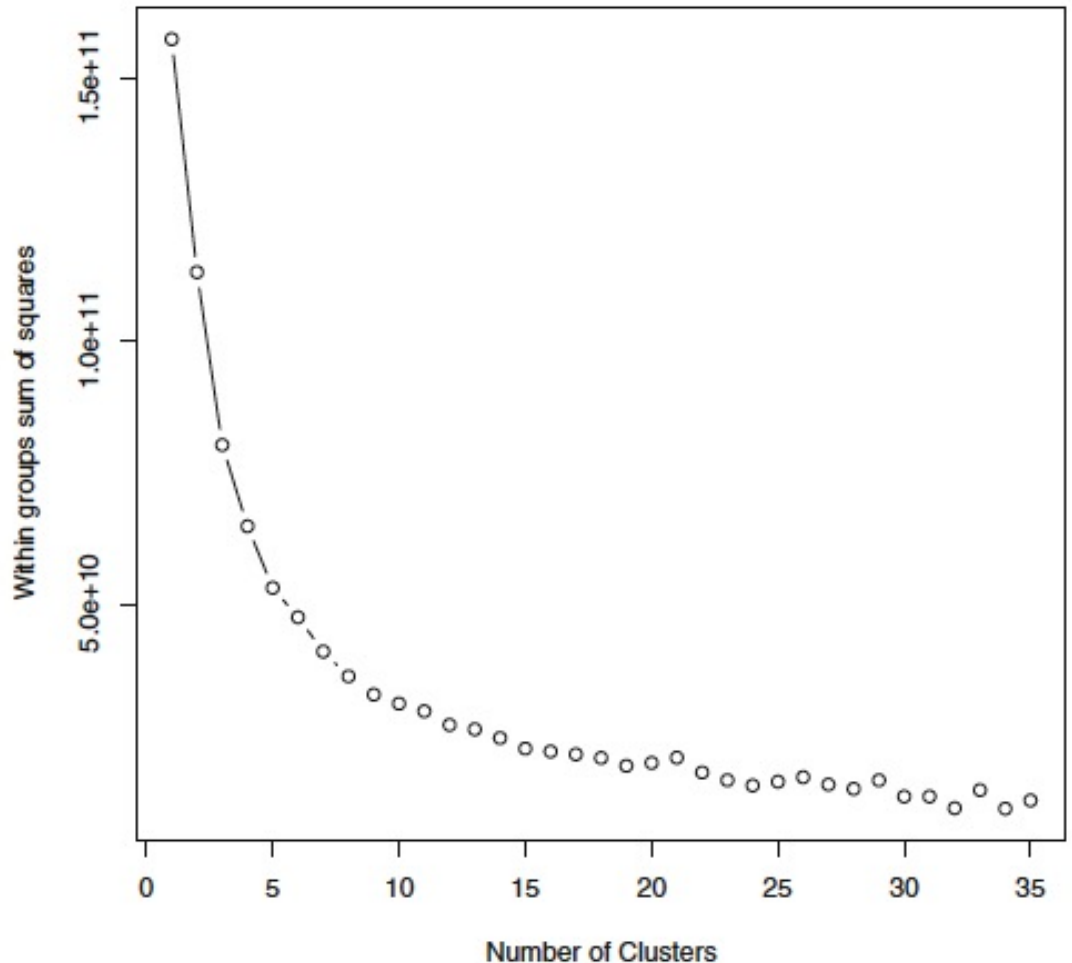
- ✱ Visualize the data with scatter plots
- ✱ We do see that some features are correlated.
- ✱ But overall we do not see significant structure or groups in the data.



Scatter Plot Matrix

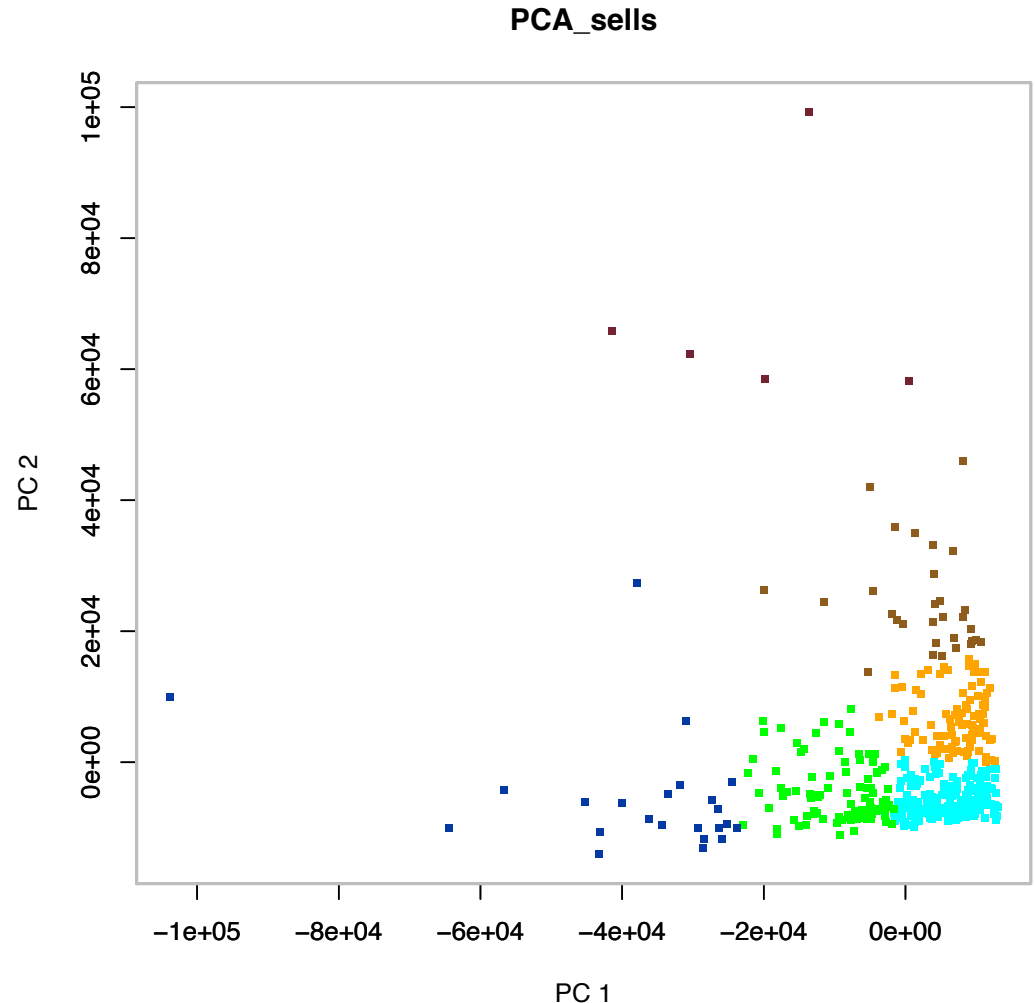
Do kmeans and choose k through the cost function

It's good to pick a **k** around the knee:
I choose 6 for it matches the number of labels



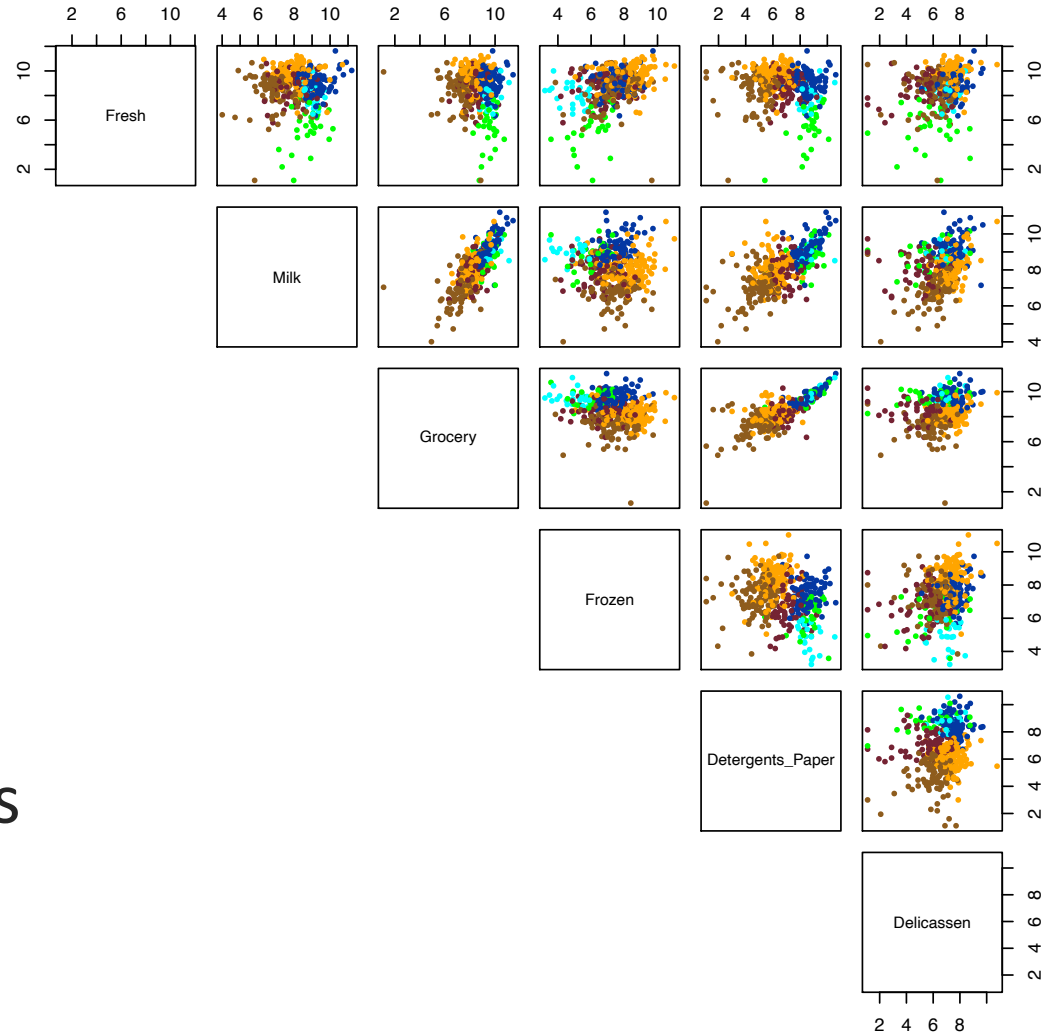
Visualization of the data (PCA)

- ✱ PCA does show some separation. **Colors are the clusters**
- ✱ Data points show large range of dynamics!



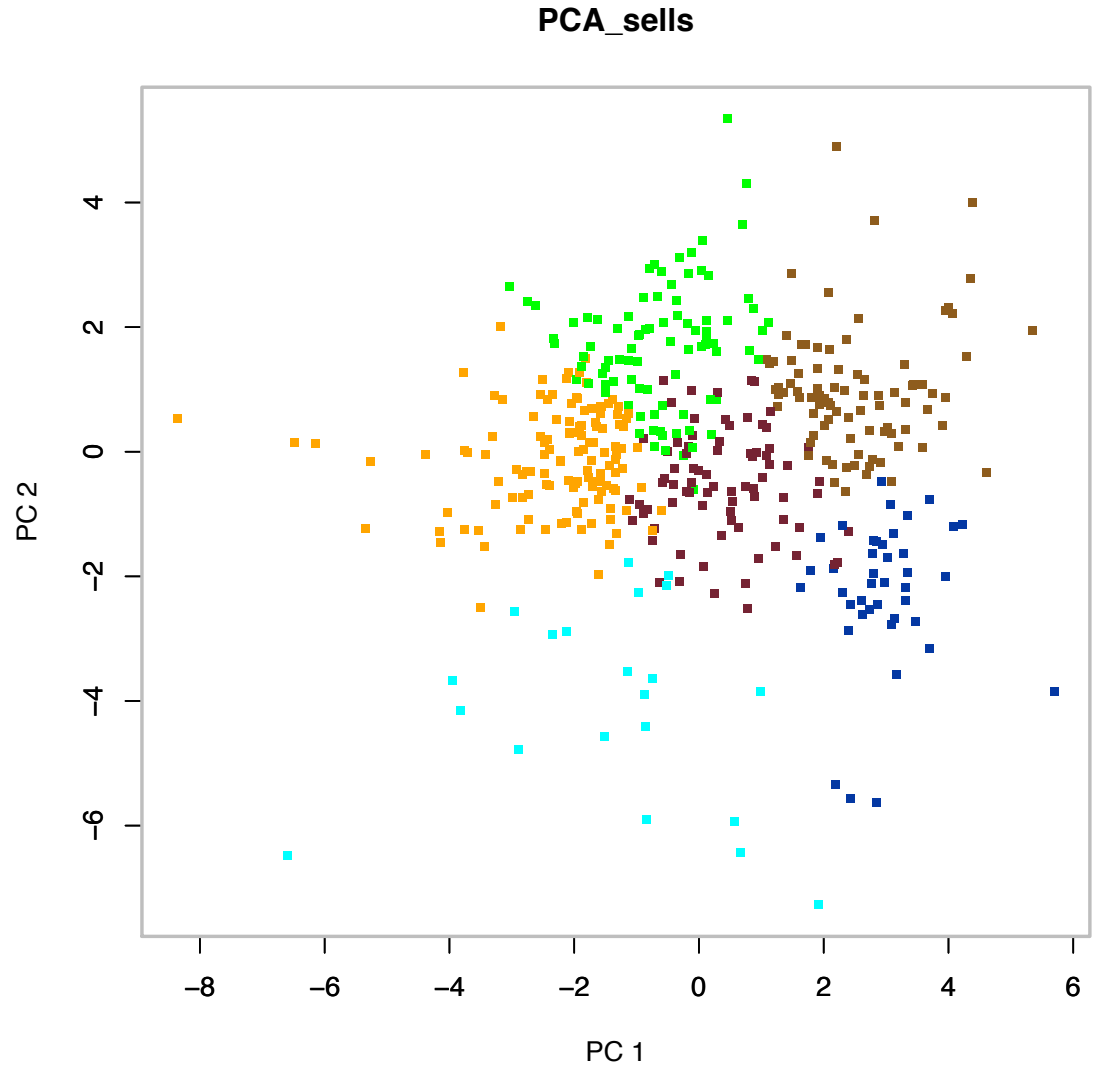
Do log transform of the data

- ✱ Log transform the data
- ✱ Do scatter plot matrix after the log transform
- ✱ Do the kmeans and color the clusters identified by k-means



PCA after log transformation: Clusters

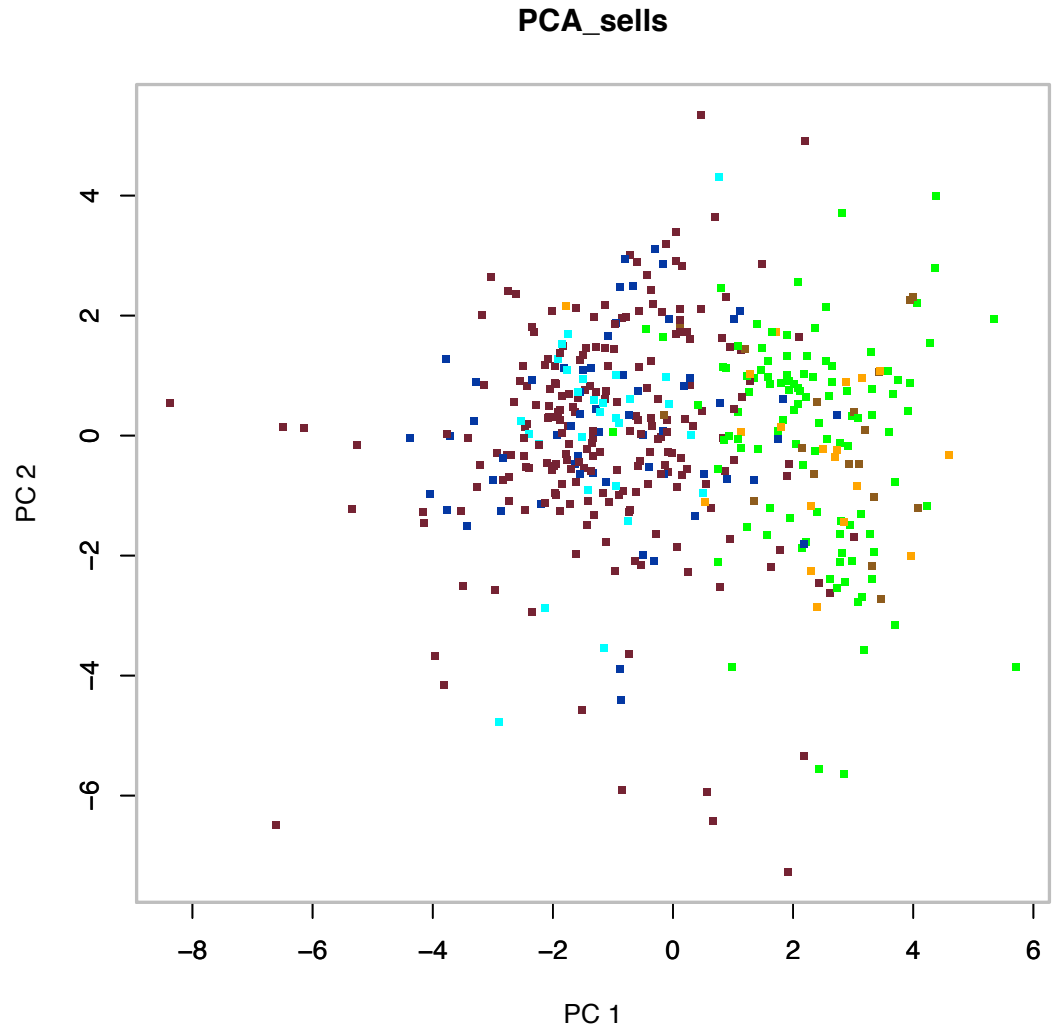
Colors show the
clusters
identified by k-
means



PCA after log transformation

Colors show the
Channel-region
labels

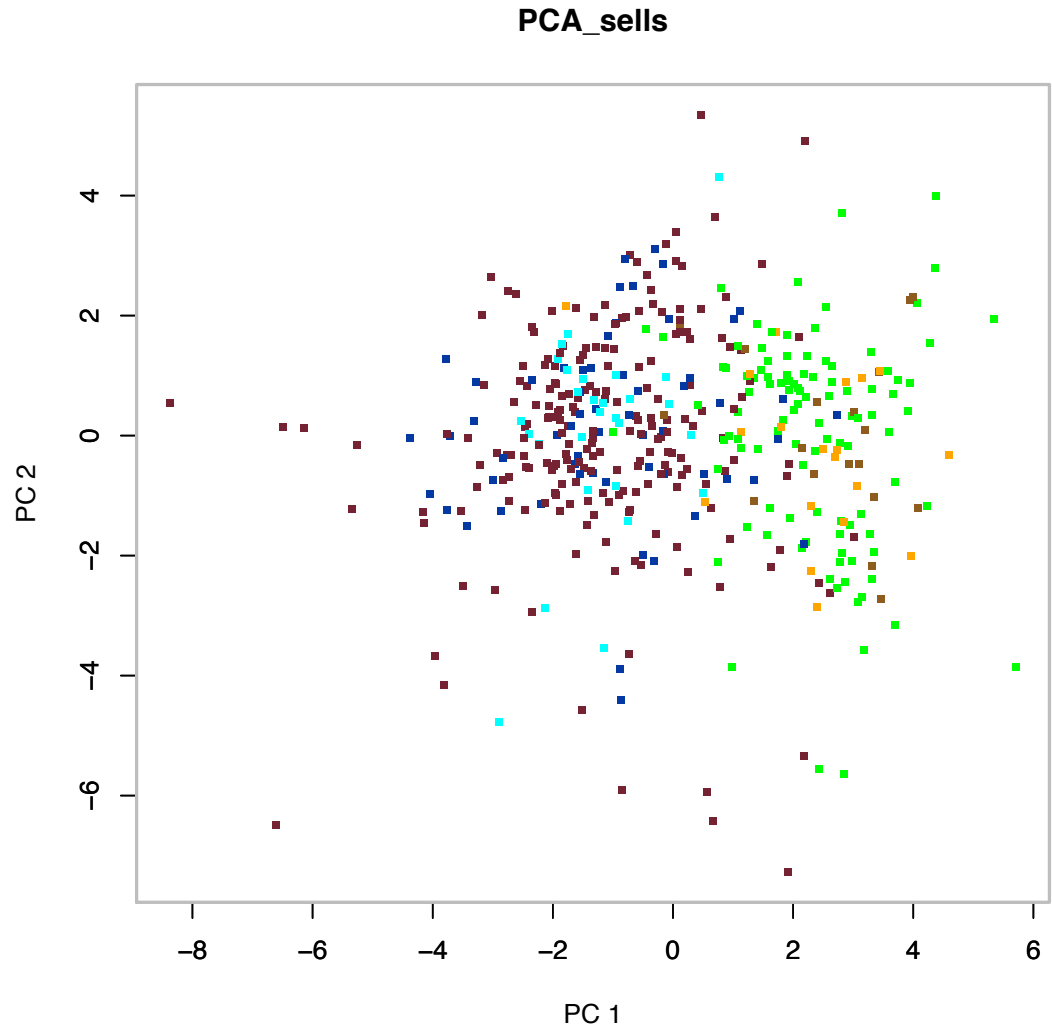
What does this
tell us?



PCA after log transformation

Colors show the
Channel-region
labels

Channels differ a
lot



Vector Quantization



Cluster center histogram of the Portugal grocery spending data

- For each channel/region, we make a histogram of customers that map to each of the **6 cluster centers**.

- What do you see?

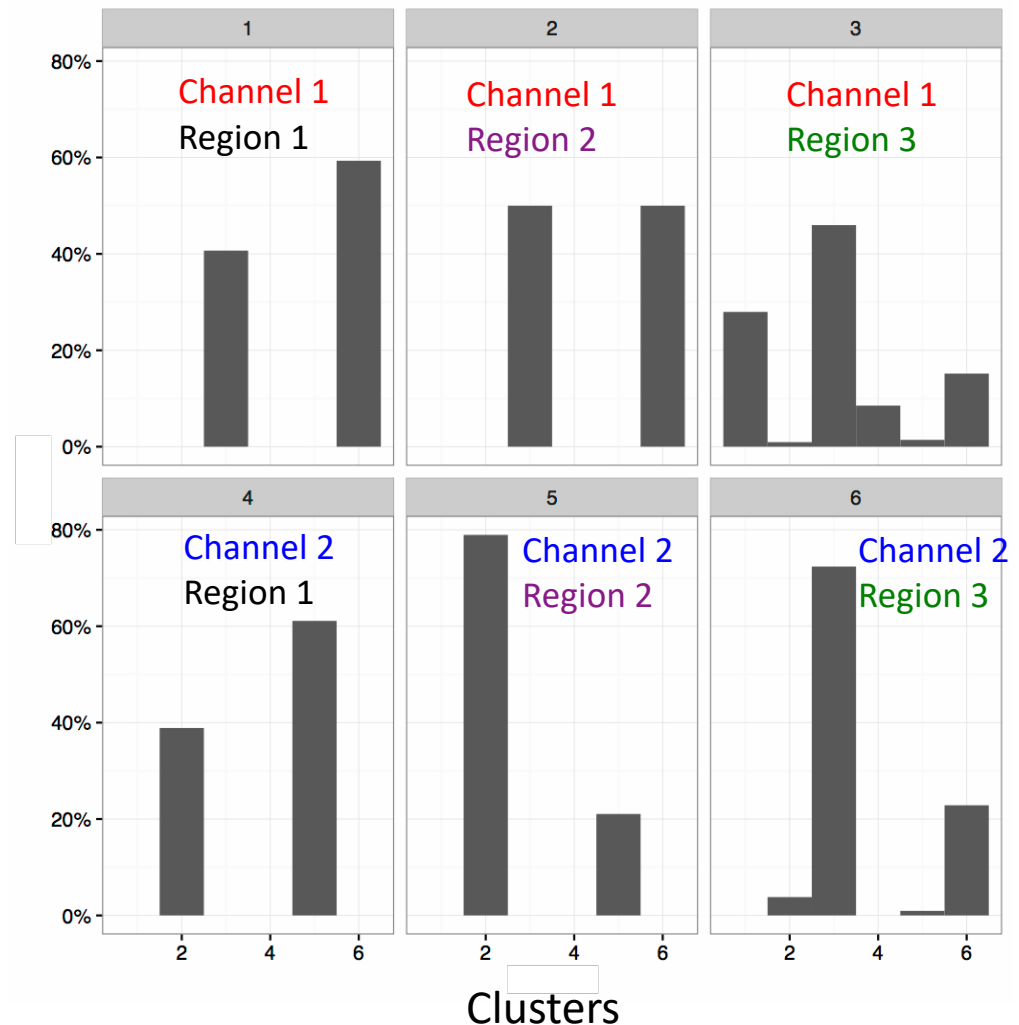
Channel1: Horeca

Channel2: Retail

Region1: Lisbon

Region2: Oporto

Region3: Other



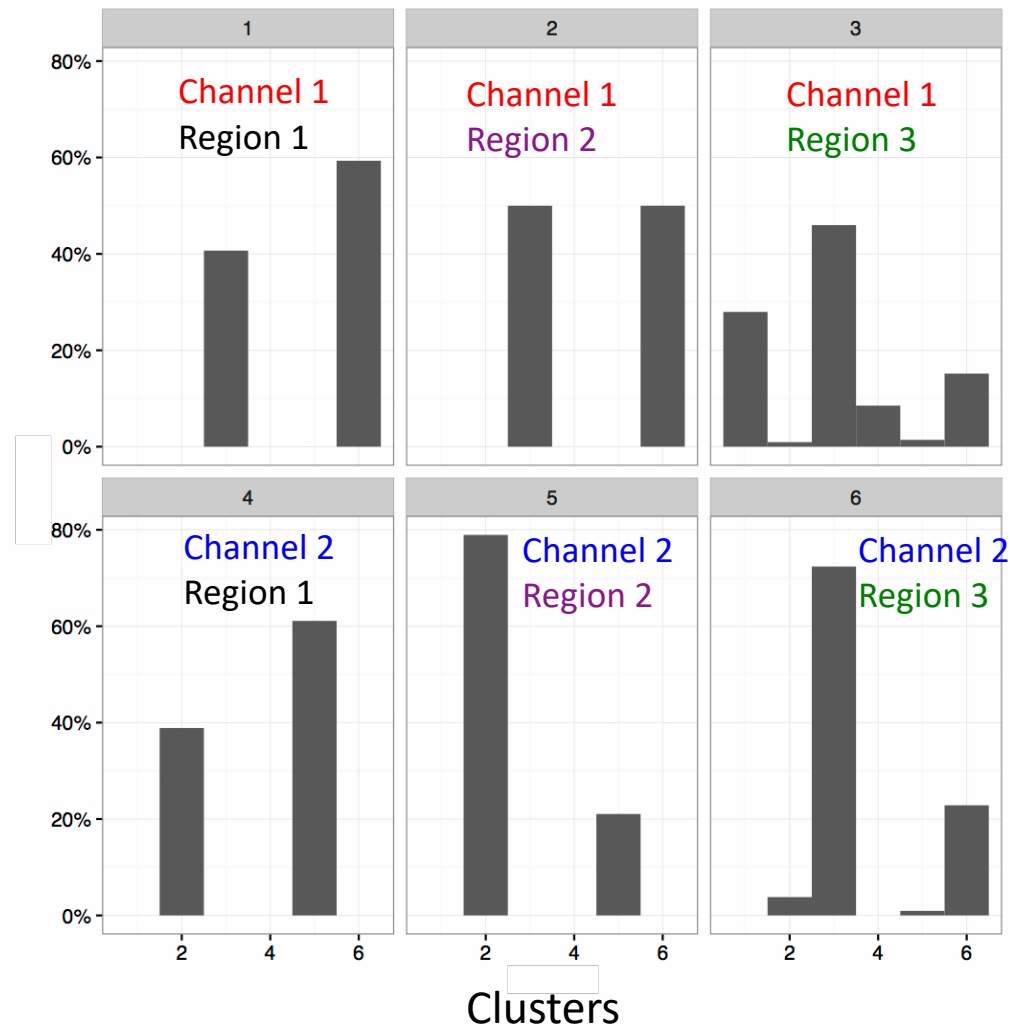
Cluster center histogram of the Portugal grocery spending data

- For each channel/region, we make a histogram of customers that map to each of the 6 cluster centers.

- Channels are significantly different!

- Region 3 is special

- Is it enough to plot the percentage?



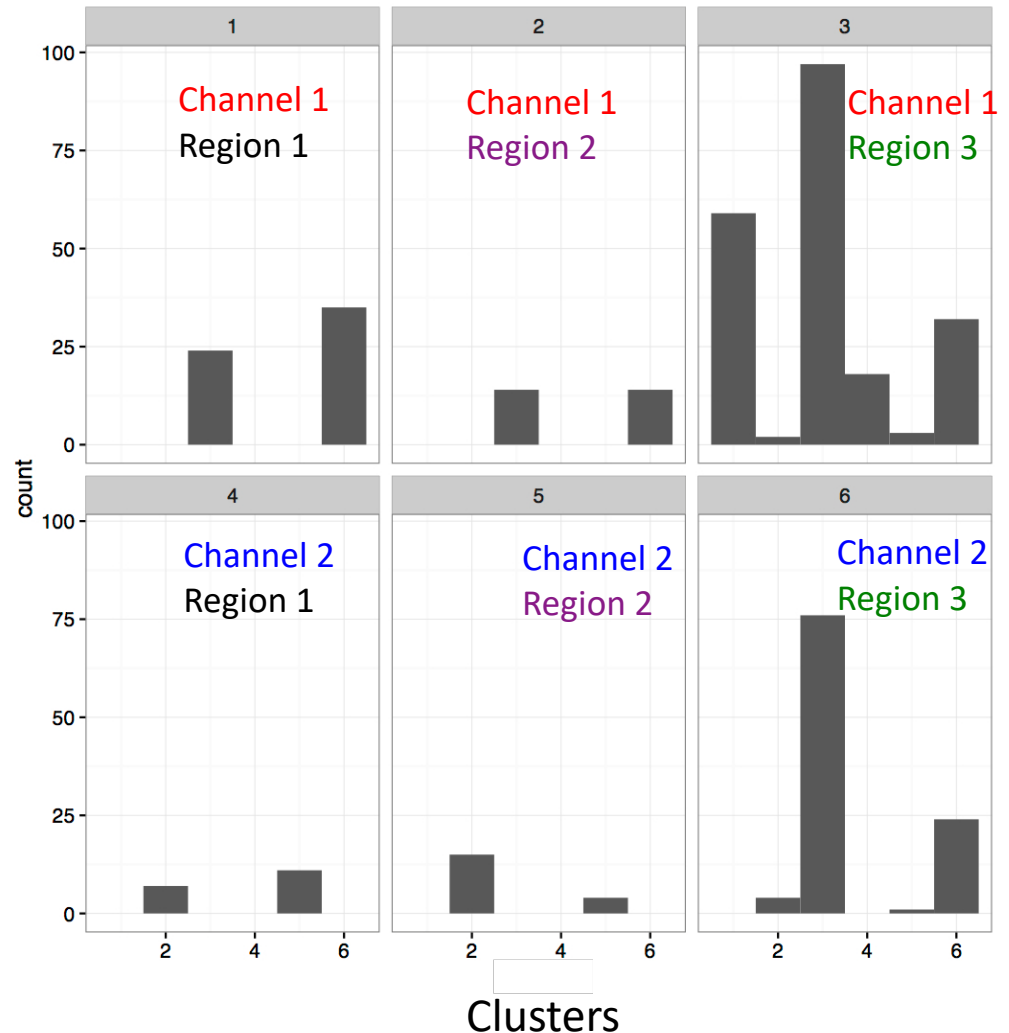
Cluster center histogram of the Portugal grocery spending data

- ✱ For each channel/region, we make a histogram of customers that map to each of the 6 cluster centers.

- ✱ Channels are significantly different!

- ✱ Region 3 is special

- ✱ Count matters depending on the purpose



Q. What can we do with cluster center histograms?

- A. investigate the feature patterns of data groups
- B. Classify new data with the cluster center histograms.
- C. Both A and B.

Vector Quantization for classifying data of varying size

- ✱ The classifiers usually assume that each feature vector has the same number of entries.
- ✱ Many datasets in fact have items of different size
 - ✱ Images usually have different numbers of pixels
 - ✱ Audio signals (and other time series) usually have different durations
- ✱ We will use **vector quantization** to map variable length data to fixed-length feature vectors using **cluster center histogram**.

Pattern vocabulary: conceptual example

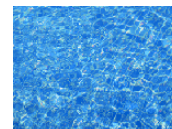
- ✱ Suppose we want to classify images into beach or prairie
- ✱ We can slice each images into 10 by 10 subsets (data entry of length 100)
- ✱ Then cluster the pieces, use the cluster center histograms to train and classify



Sand



Water



Grass

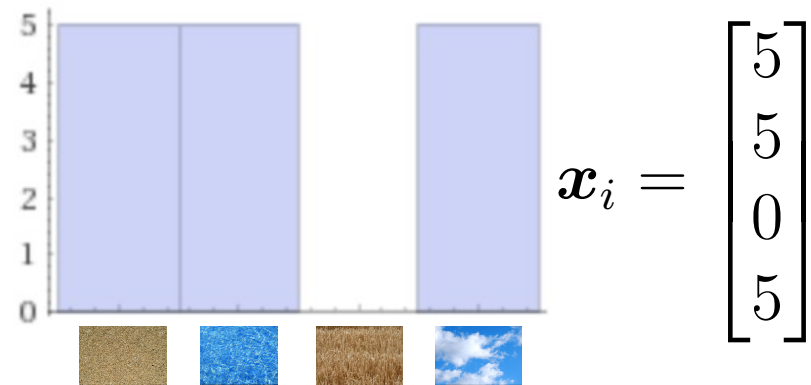
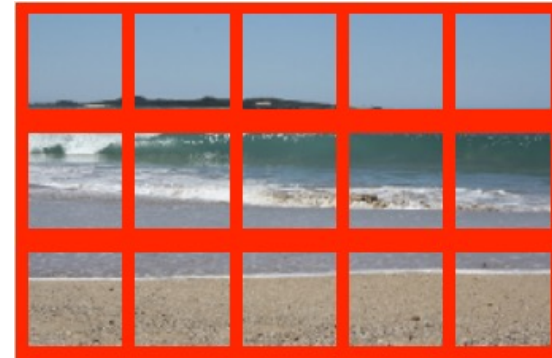


Sky



Generate fixed-length feature vectors : conceptual example

- ✱ Slice the images into 10 by 10 pixel subsets
- ✱ Do clustering on all the subsets from the training images
- ✱ Assign each subset to the nearest cluster centers (in k clusters/patterns)
- ✱ For each image, produce the counts with respect to each cluster center and form a feature vector of dimension k



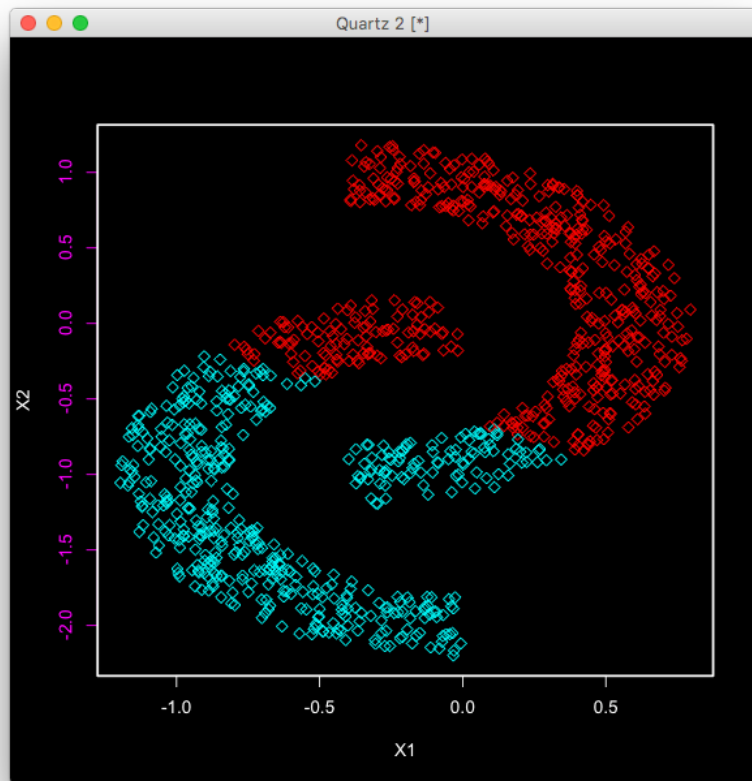
Spectral clustering



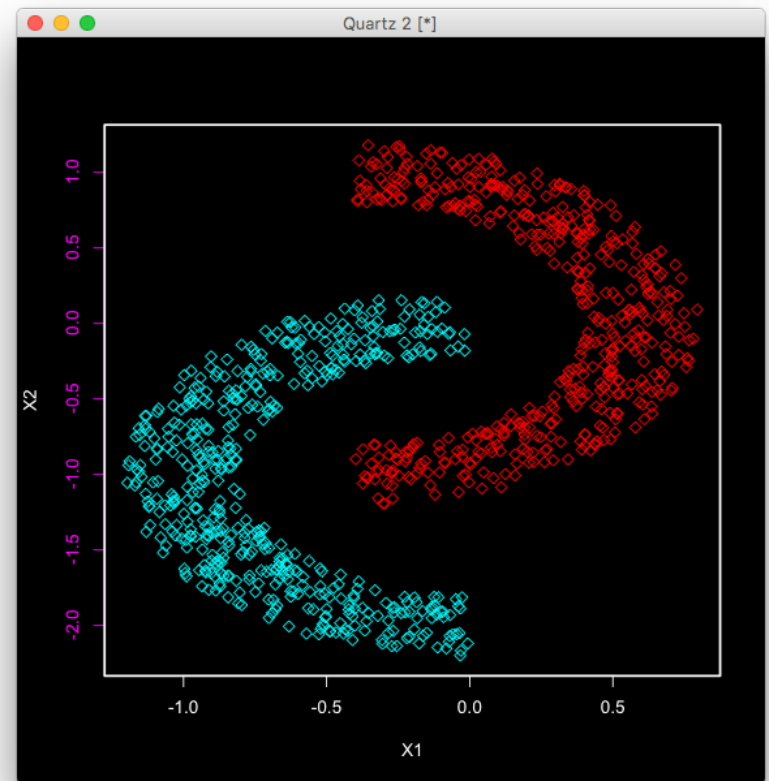
K-means is limited

✱ K-means fails in the **Two-moon** problem

K-means result



Expected result



Spectral Clustering

✱ Theoretical basis

- ✱ The Graph Representation
- ✱ The Adjacency Matrix
- ✱ Graph cut
- ✱ The Laplacian Matrix
- ✱ The properties of Laplacian that point to the solution

Again it's about Matrix !



Spectral Clustering

✱ Theoretical basis

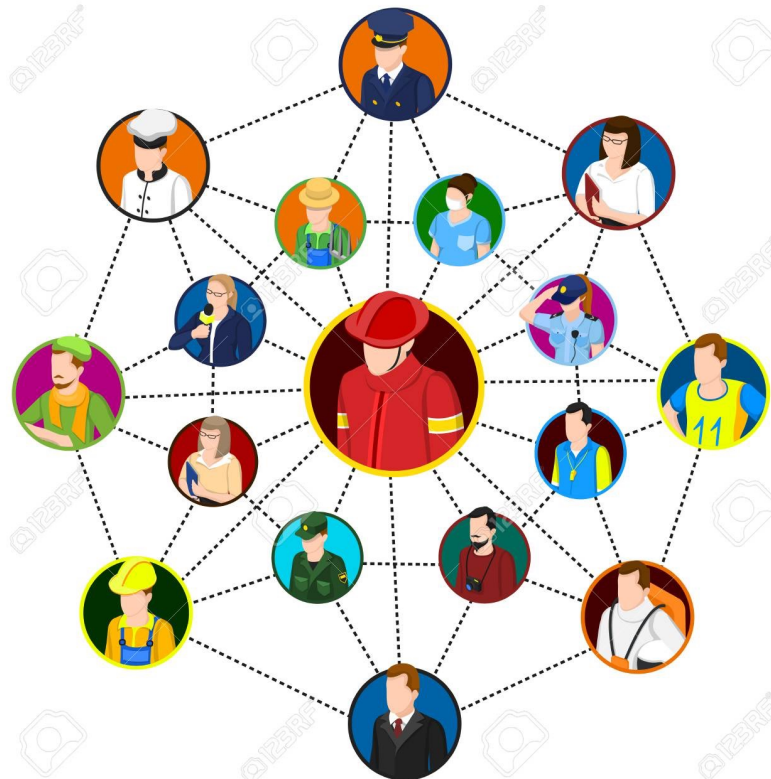
✱ *The Graph Representation*

Introduction of Graph

✱ Real world data often needs graph

✱ Strength

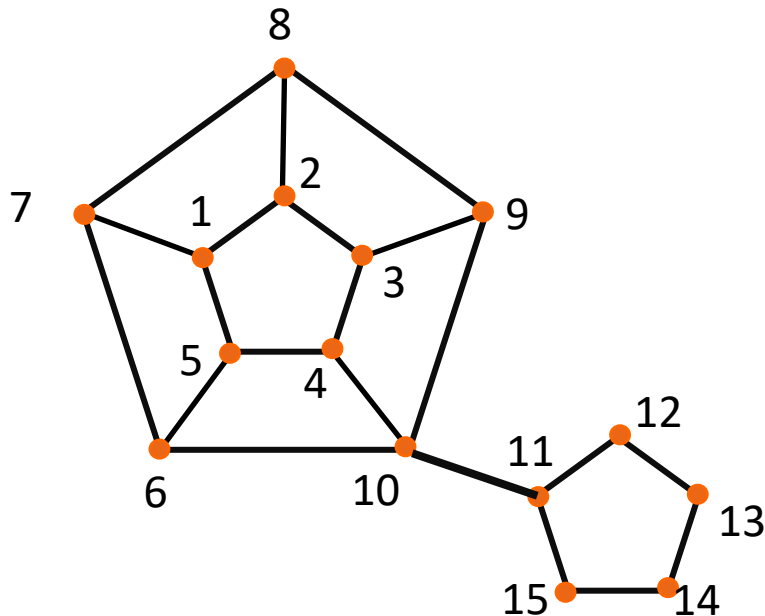
*A graph model of
Social network data*



Graph in terms of Mathematics

✱ The graph is a set $G(V, E)$

✱ V is the set of vertices



15 vertices, 21 edges

✱ E is the set of edges, showing the relationship between pair of vertices

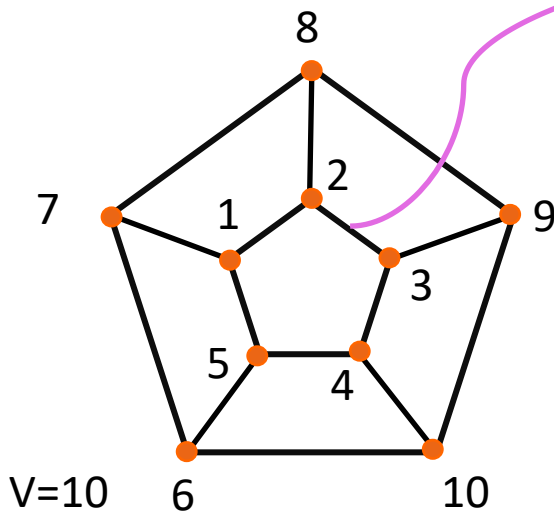
Spectral Clustering

✱ Theoretical basis

- ✱ The Graph Representation
- ✱ *The Adjacency Matrix*

Graph data in the format of matrix

These 10 geometric data points can be represented with an *undirected* Graph and then numerically written as a matrix



V=10
E=15

N=10

Adjacency Matrix*: W $\{\omega_{ij} \geq 0\}$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	1	0	1	0	0	0
2	1	0	1	0	0	0	0	1	0	0
3	0	1	0	1	0	0	0	0	1	0
4	0	0	1	0	1	0	0	0	0	1
5	1	0	0	1	0	1	0	0	0	0
6	0	0	0	0	1	0	1	0	0	1
7	1	0	0	0	0	1	0	1	0	0
8	0	1	0	0	0	0	1	0	1	0
9	0	0	1	0	0	0	0	1	0	1
10	0	0	0	1	0	1	0	0	1	0

* Some people prefer "Similarity matrix"

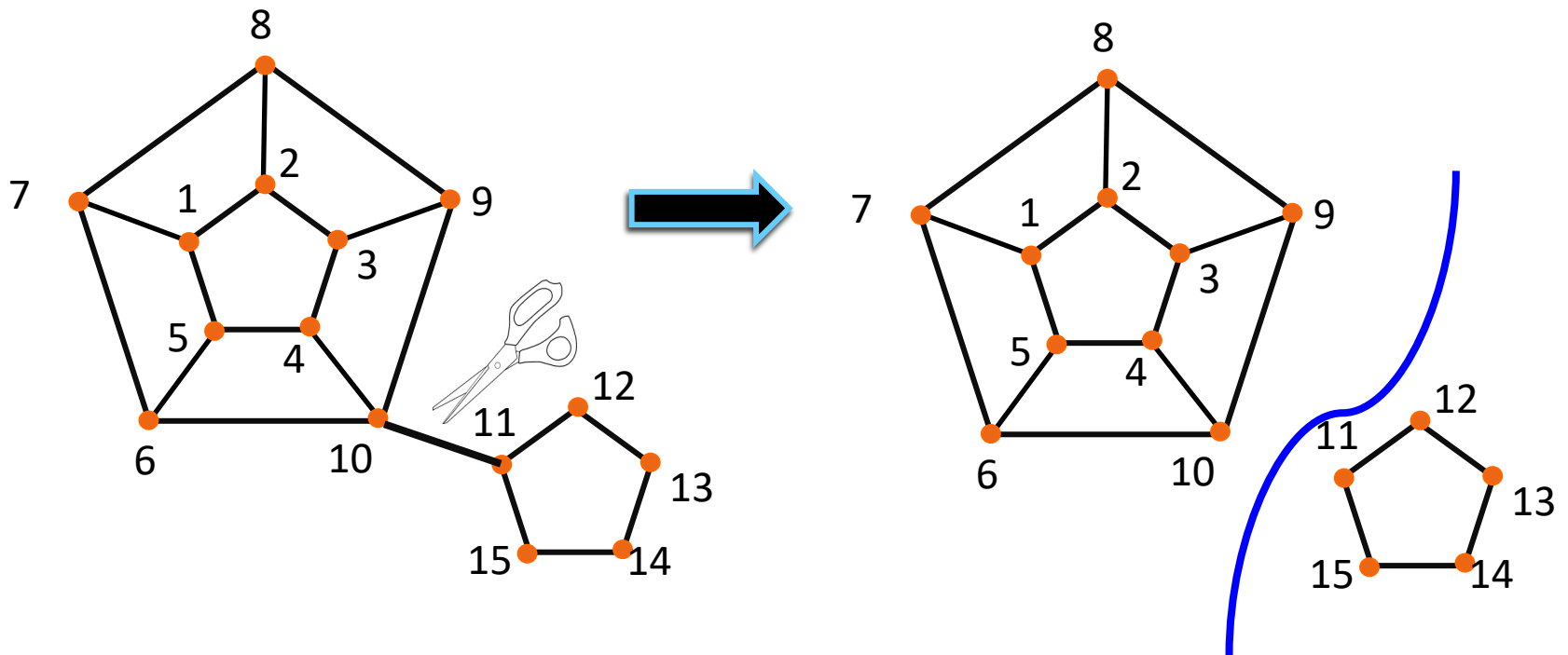
Spectral Clustering

✱ Theoretical basis

- ✱ The Graph Representation
- ✱ The Adjacency Matrix
- ✱ ***Graph cut***

Spectral Clustering emerged from Graph-cut

✻ Clusters are learned via min-Cut of the Graph



Spectral Clustering vs Graph-cut

✱ Spectral clustering is equivalent to the Graph-cut

Finding clusters is to solve an **Eigenvalue problem** using **Graph's Laplacian matrix**

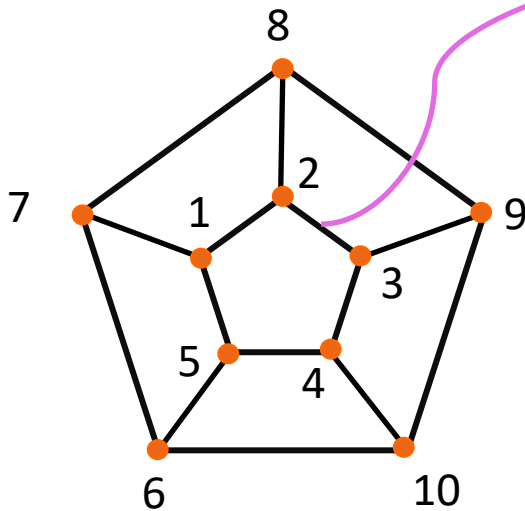
Spectral Clustering

✱ Theoretical basis

- ✱ The Graph Representation
- ✱ The Adjacency Matrix
- ✱ Graph cut
- ✱ ***The Laplacian Matrix***

Graph data in the format of matrix

The weights ω_{ij} of the edges stored in the matrix of the graph can be **any non-negative values**



Adjacency Matrix: W $\{\omega_{ij} \geq 0\}$

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	0	1	0	1	0	0	0
2	1	0	1	0	0	0	0	1	0	0
3	0	1	0	1	0	0	0	0	1	0
4	0	0	1	0	1	0	0	0	0	1
5	1	0	0	1	0	1	0	0	0	0
6	0	0	0	0	1	0	1	0	0	1
7	1	0	0	0	0	1	0	1	0	0
8	0	1	0	0	0	0	1	0	1	0
9	0	0	1	0	0	0	0	1	0	1
10	0	0	0	1	0	1	0	0	1	0

Transform Adjacency Matrix **W** into Graph Laplacian Matrix **L**

$$\text{⊗} \quad \mathbf{L} = \mathbf{D} - \mathbf{W}$$

$$\mathbf{D}_{ij} = \begin{cases} \sum_k \omega_{ik}, & i=j \\ 0, & i \neq j \end{cases}$$

Adjacency Matrix: **W** $\{\omega_{ij}\}$

	1	2	3	4	5	6	7	8	9	X
1	0	1	0	0	1	0	1	0	0	0
2	1	0	1	0	0	0	0	1	0	0
3	0	1	0	1	0	0	0	0	1	0
4	0	0	1	0	1	0	0	0	0	1
5	1	0	0	1	0	1	0	0	0	0
6	0	0	0	0	1	0	1	0	0	1
7	1	0	0	0	0	1	0	1	0	0
8	0	1	0	0	0	0	1	0	1	0
9	0	0	1	0	0	0	0	1	0	1
X	0	0	0	1	0	1	0	0	1	0



Laplacian Matrix: **L** $\{L_{ij}\}$

	1	2	3	4	5	6	7	8	9	X
1	3	-1	0	0	-1	0	-1	0	0	0
2	-1	3	-1	0	0	0	0	-1	0	0
3	0	-1	3	-1	0	0	0	0	-1	0
4	0	0	-1	3	-1	0	0	0	0	-1
5	-1	0	0	-1	3	-1	0	0	0	0
6	0	0	0	0	-1	3	-1	0	0	-1
7	-1	0	0	0	0	-1	3	-1	0	0
8	0	-1	0	0	0	0	-1	3	-1	0
9	0	0	-1	0	0	0	0	-1	3	-1
X	0	0	0	-1	0	-1	0	0	-1	3

Q. What properties do you see in L matrix?

Adjacency Matrix: $\mathbf{W} \{\omega_{ij}\}$

	1	2	3	4	5	6	7	8	9	X
1	0	1	0	0	1	0	1	0	0	0
2	1	0	1	0	0	0	0	1	0	0
3	0	1	0	1	0	0	0	0	1	0
4	0	0	1	0	1	0	0	0	0	1
5	1	0	0	1	0	1	0	0	0	0
6	0	0	0	0	1	0	1	0	0	1
7	1	0	0	0	0	1	0	1	0	0
8	0	1	0	0	0	0	1	0	1	0
9	0	0	1	0	0	0	0	1	0	1
X	0	0	0	1	0	1	0	0	1	0



Laplacian Matrix: $\mathbf{L} \{L_{ij}\}$

	1	2	3	4	5	6	7	8	9	X
1	3	-1	0	0	-1	0	-1	0	0	0
2	-1	3	-1	0	0	0	0	-1	0	0
3	0	-1	3	-1	0	0	0	0	-1	0
4	0	0	-1	3	-1	0	0	0	0	-1
5	-1	0	0	-1	3	-1	0	0	0	0
6	0	0	0	0	-1	3	-1	0	0	-1
7	-1	0	0	0	0	-1	3	-1	0	0
8	0	-1	0	0	0	0	-1	3	-1	0
9	0	0	-1	0	0	0	0	-1	3	-1
X	0	0	0	-1	0	-1	0	0	-1	3

Spectral Clustering

✱ Theoretical basis

- ✱ The Graph Representation
- ✱ The Adjacency Matrix
- ✱ Graph cut
- ✱ The Laplacian Matrix
- ✱ ***The properties of Laplacian that point to the solution***

Laplacian Matrix L 's properties

$$\ast L = D - W$$

$$D_{ij} = \begin{cases} \sum_k \omega_{ik}, & i=j \\ 0, & i \neq j \end{cases}$$

Laplacian Matrix: L ($\{L_{ij}\}$)

	1	2	3	4	5	6	7	8	9	X
1	3	-1	0	0	-1	0	-1	0	0	0
2	-1	3	-1	0	0	0	0	-1	0	0
3	0	-1	3	-1	0	0	0	0	-1	0
4	0	0	-1	3	-1	0	0	0	0	-1
5	-1	0	0	-1	3	-1	0	0	0	0
6	0	0	0	0	-1	3	-1	0	0	-1
7	-1	0	0	0	0	-1	3	-1	0	0
8	0	-1	0	0	0	0	-1	3	-1	0
9	0	0	-1	0	0	0	0	-1	3	-1
X	0	0	0	-1	0	-1	0	0	-1	3

Properties (I—III)

(I) Symmetric

(II) Row Sums = 0

(III) Quadratic form

$$f' L f = \frac{1}{2} \sum_{ij} \omega_{ij} (f_i - f_j)^2 \geq 0$$

f is any nonzero vector

Energy function

Laplacian Matrix L 's properties (p4)

$$\ast \mathbf{L} = \mathbf{D} - \mathbf{W}$$

$$\mathbf{D}_{ij} = \begin{cases} \sum_k \omega_{ik}, & i=j \\ 0, & i \neq j \end{cases}$$

$$\ast \mathbf{L} \mathbf{x} = \lambda \mathbf{x}$$

Property (IV):

Positive semi-definite

All $\lambda_i \geq 0$ and there is at least one eigenvalue

$\lambda_0 = 0$ s.t. $\mathbf{u}_0 = \{1, 1 \dots 1\}$ constant vector

$\underbrace{\hspace{1.5cm}}_n$

Laplacian Matrix **L**'s properties (p5)

$$\ast \mathbf{L} = \mathbf{D} - \mathbf{W}$$

$$\mathbf{D}_{ij} = \begin{cases} \sum_k \omega_{ik}, & i=j \\ 0, & i \neq j \end{cases}$$

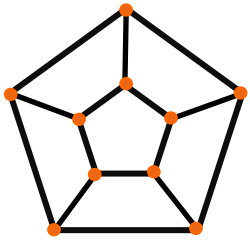
$$\ast \mathbf{L} \mathbf{x} = \lambda \mathbf{x}$$

Property (V):

of zero valued λ_i is equal to the number of clusters (connected subgraph) in the graph

Eigenvalue distributions of three examples

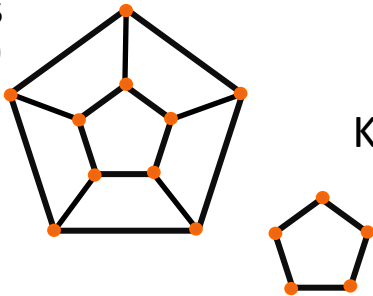
V=10
E=15



A

K=1

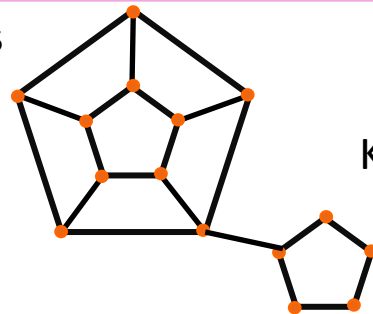
V=15
E=20



B

K=2

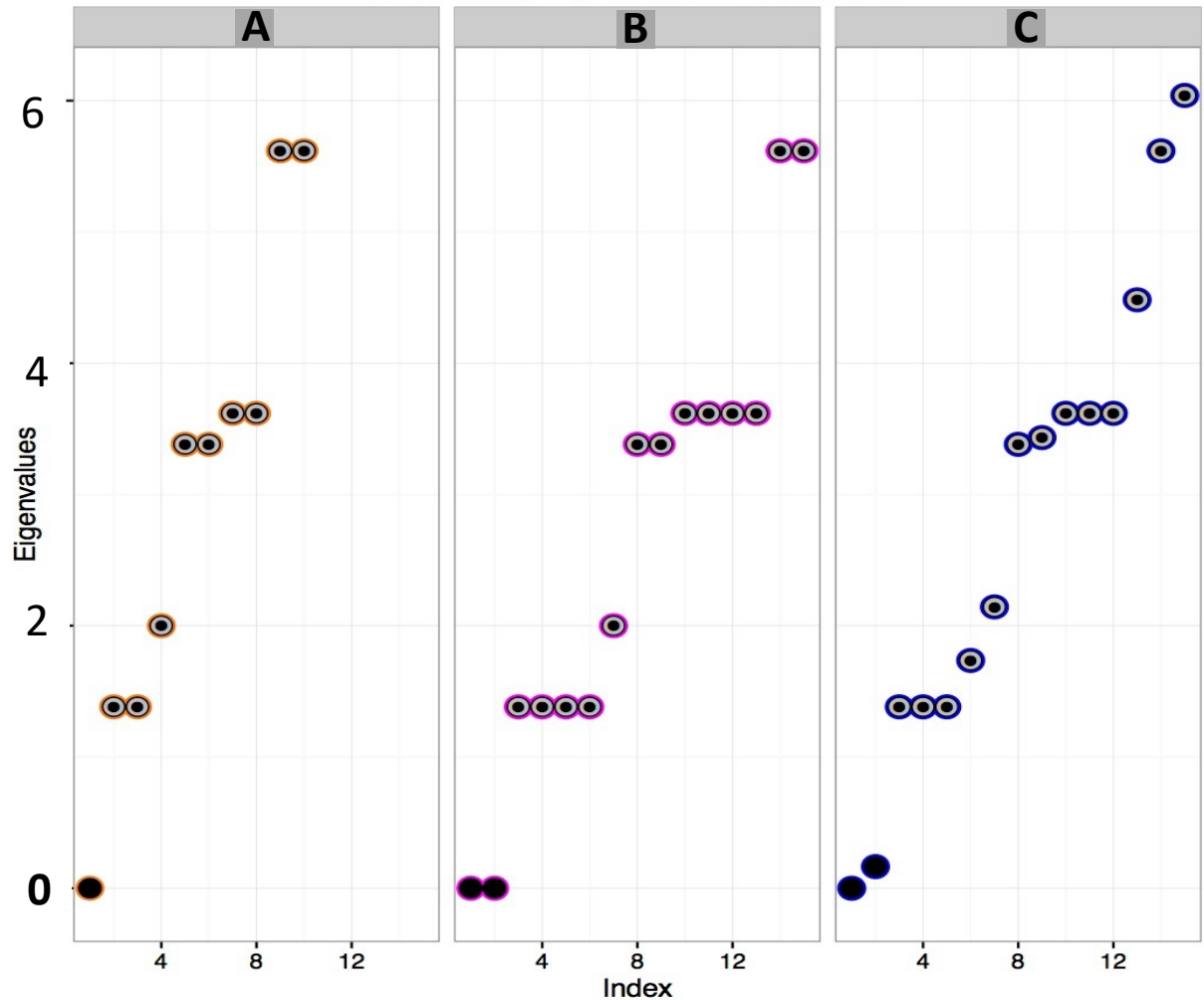
V=15
E=21



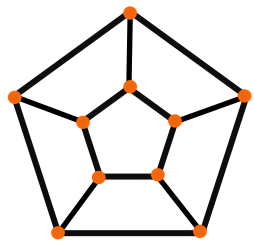
C

K=?

Ascending Eigenvalues Distribution

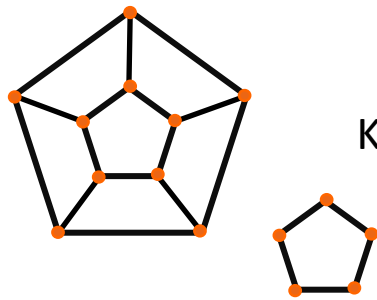


Eigenvalue distributions of three examples



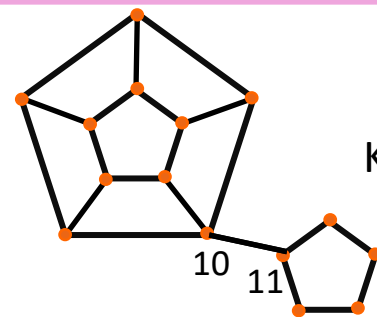
A

K=1



B

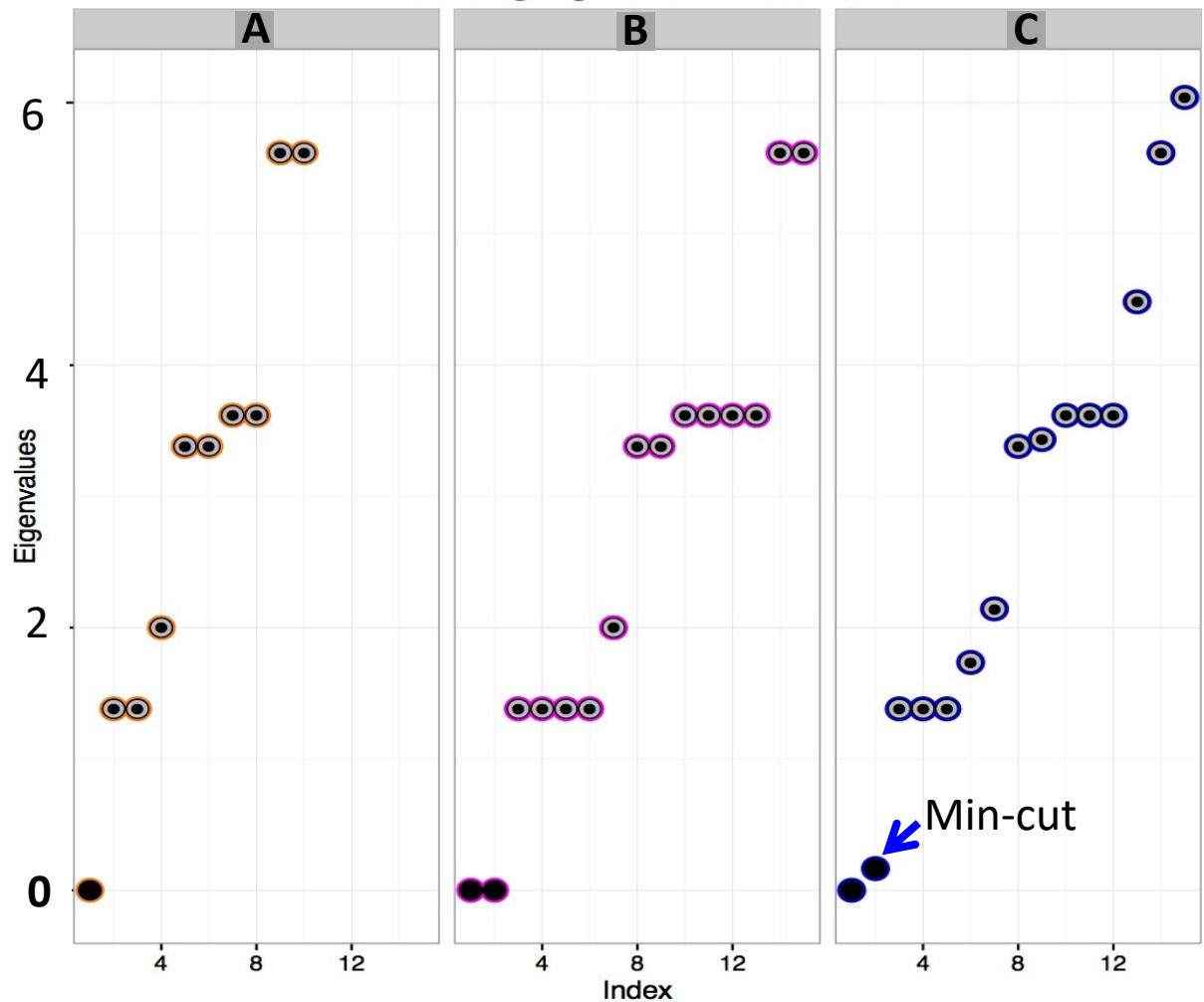
K=2



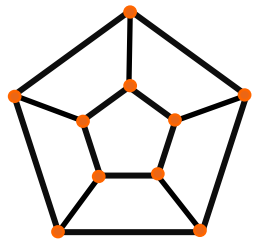
C

K=1

Ascending Eigenvalues Distribution

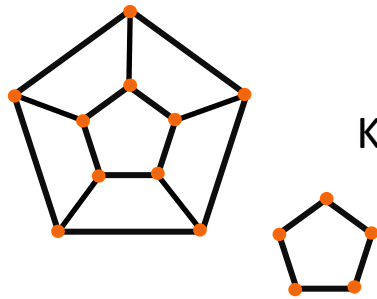


First two Eigenvectors of three examples



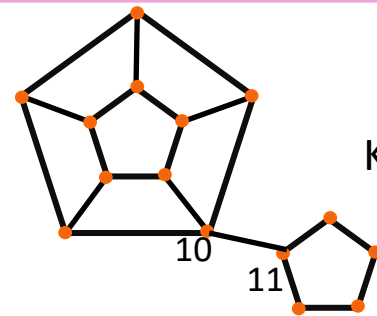
A

K=1



B

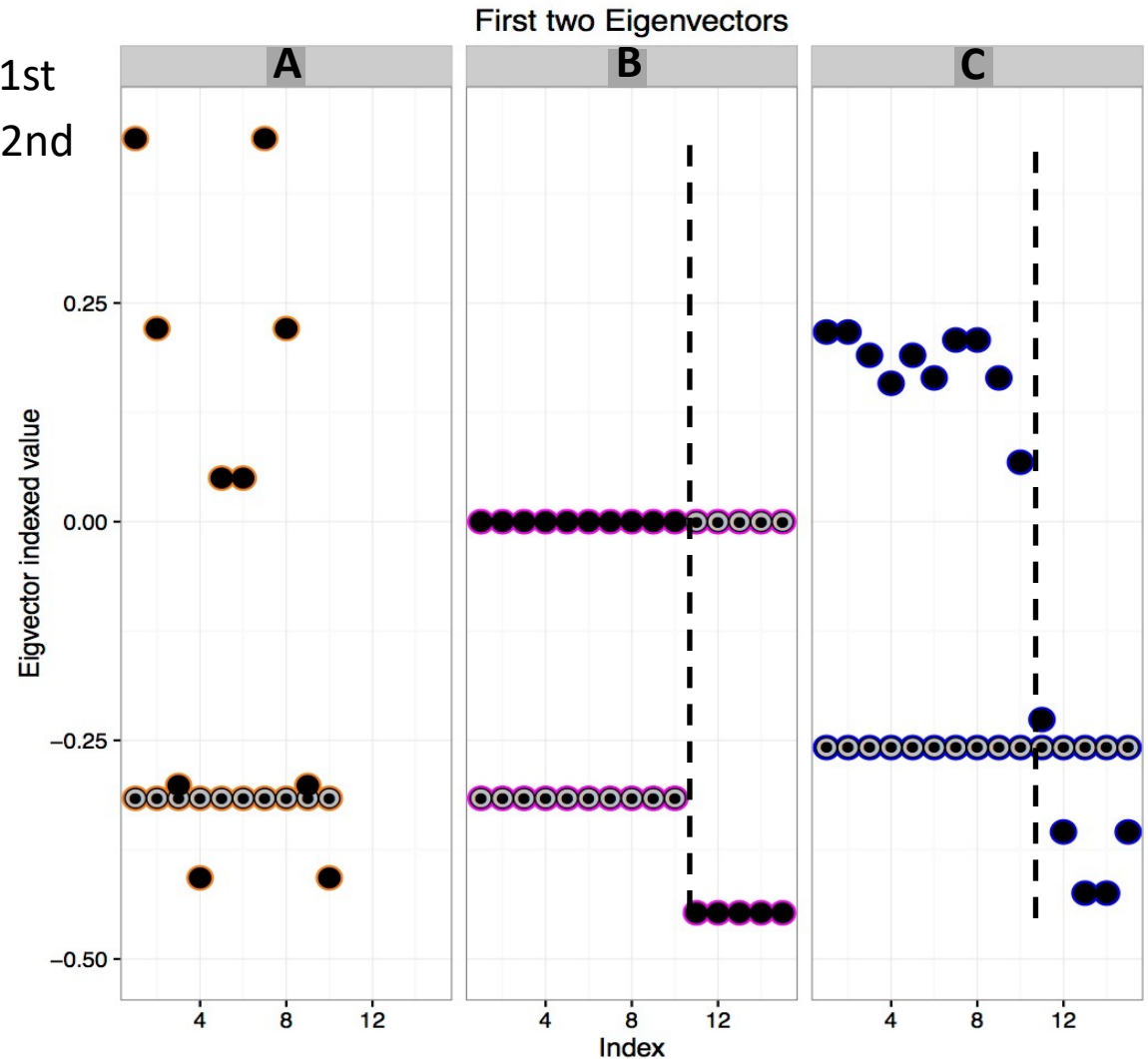
K=2



C

K=1

- 1st
- 2nd



Discussion

- ✱ Why does Spectral Clustering perform better than k-means for non-convex shaped data?

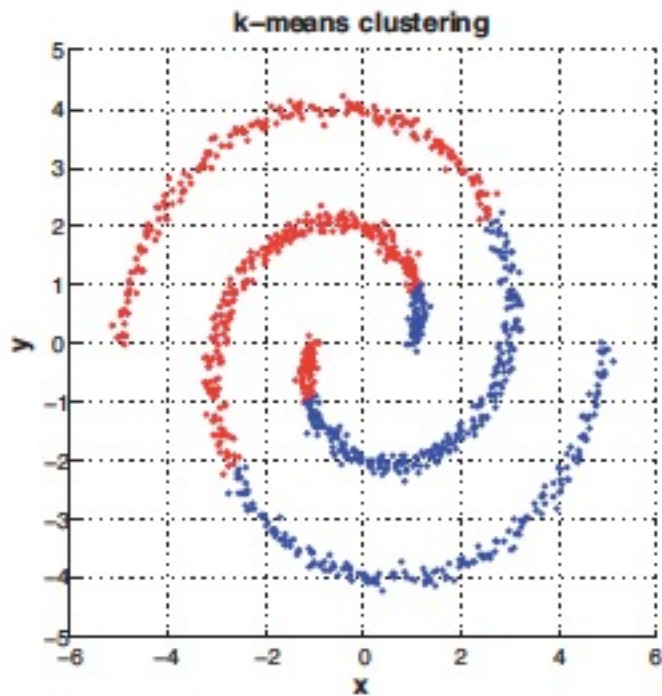
Discussion

- ✱ Why does Spectral Clustering perform better than k-means for non-convex shaped data?
 - i) Graph representation kept the topological relationship btw data

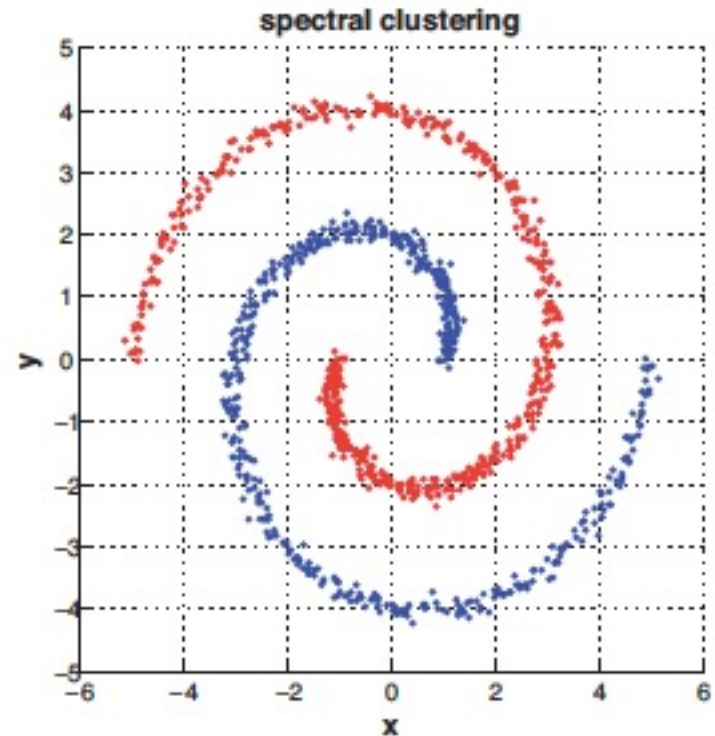
Discussion

- ✱ Why does Spectral Clustering perform better than k-means for non-convex shaped data?
 - i) Graph representation kept the topological relationship btw data
 - ii) Eigenvectors are piecewise constant in the ideal cases, which are easy to cluster

Some Spectral clustering results



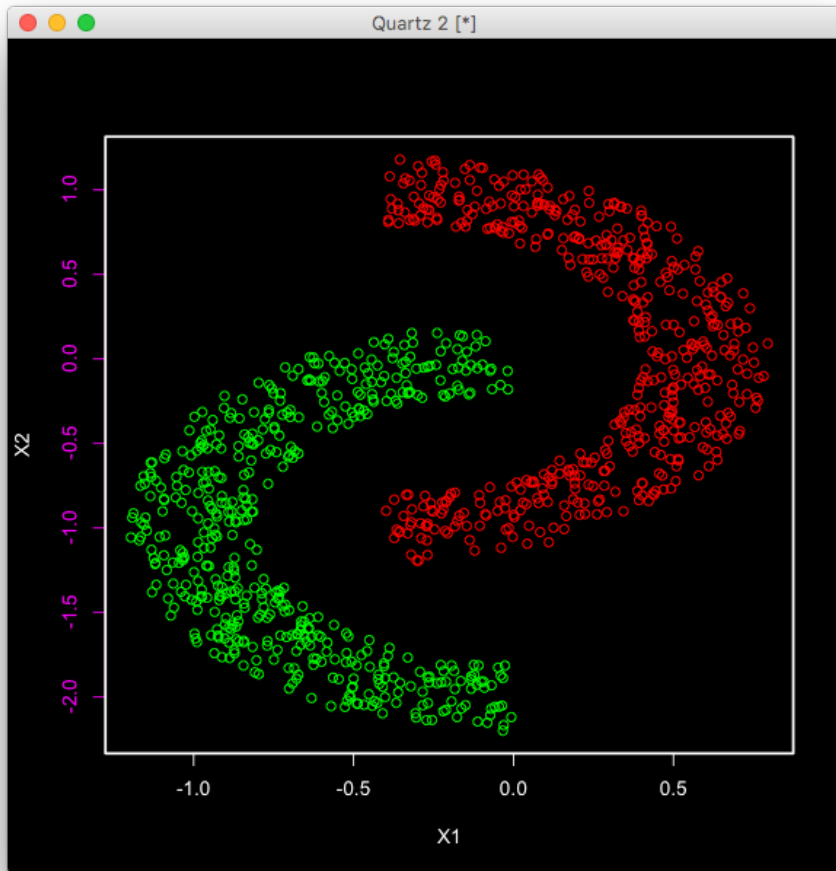
(a)



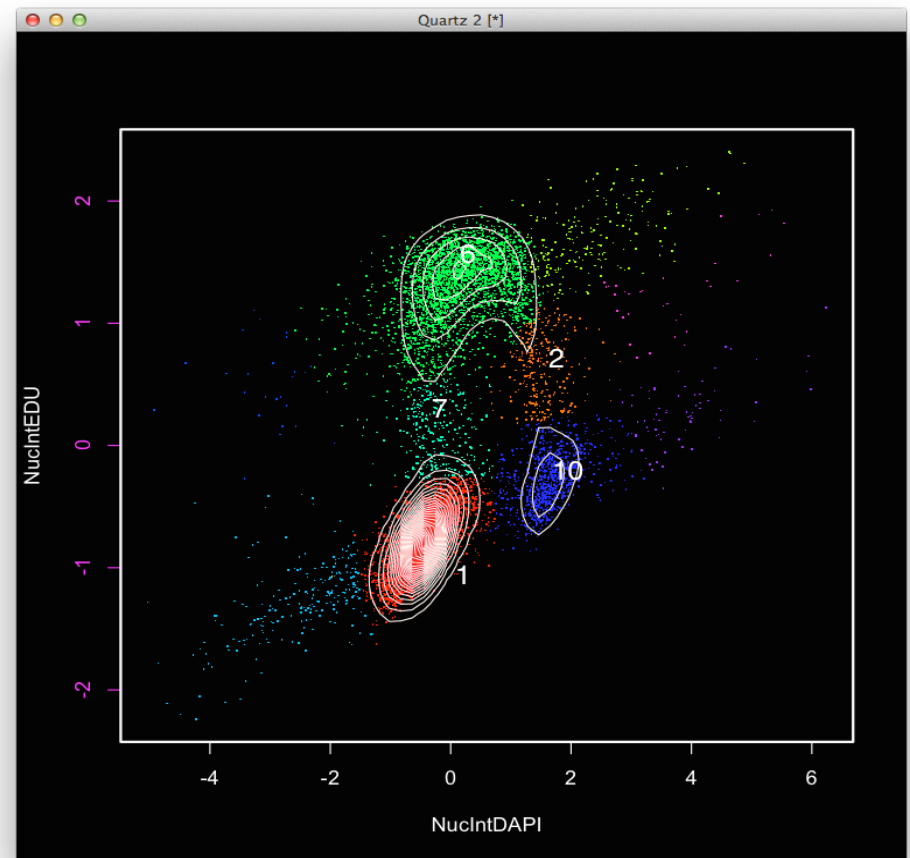
(b)

Some Spectral clustering results

Two-Moons



Cell Cycle phases



Conclusion of Spectral Clustering

Given # of zero valued λ_i = the number of disconnected components of the graph, we can *approximately* use the first k number of eigenvectors to cluster the data into k clusters.

The intuition: The singularities of the graph's Laplacian correspond to the # of clusters in the graph.

Assignments

- ✱ Finish Chapter 12 of the textbook
- ✱ Next time: Markov chain

Additional References

- ✱ Robert V. Hogg, Elliot A. Tanis and Dale L. Zimmerman. “Probability and Statistical Inference”
- ✱ Kelvin Murphy, “Machine learning, A Probabilistic perspective”

See you next time

*See
You!*

